

محمد منيع

M78.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:406460319

Submission Date

Aug 8, 2026, 11:05 PM GMT+3

Download Date

Aug 9, 2026, 4:06 AM GMT+3

File Name

M78.docx

File Size

319.3 KB

19 Pages

2,477 Words

16,550 Characters

86% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely revised by AI.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.



15 AI-generated 86%

Likely generated or likely generated and revised by AI.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely human generated as AI generated and likely AI generated as human generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.



A Predictive Model for Student Academic Performance in Higher Education Using Advanced Machine Learning Techniques

Prepared By

Name: AbdulMoien Al Ali

ID No.: 261002308

Name: Mohammed Almania

ID No.: 261002689

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

Supervisor Name: Dr.Sherif Mostafa

Collage of Informatics, Midocean University, Comoros

Email:.....

Acknowledgements

Special thanks to Dr. .

Abstract

This research addresses the critical challenge of student attrition and academic failure in higher education by developing a robust predictive model. As educational institutions transition towards data-driven decision-making, the ability to identify at-risk students early is paramount. This study utilizes the mathematics dataset from the UCI Machine Learning Repository, encompassing academic, demographic, and social attributes. By applying advanced machine learning techniques, including Random Forest, Support Vector Machines (SVM), and Gradient Boosting, the model aims to provide high-accuracy predictions of student success. Furthermore, the integration of SHAP (SHapley Additive exPlanations) ensures model interpretability, allowing educators to understand the underlying factors influencing each prediction. The findings demonstrate that proactive intervention, guided by intelligent systems, can significantly enhance student outcomes and institutional efficiency.

List of Publications

Table of Contents

List of Abbreviations	vii
List of Figures.....	viii
List of Tables	ix

Chapter 1: Introduction and

Overview..... 1

Chapter 2: Background and Related Work

..... 31

Chapter 3: Proposed Approach 78

4.1 Introduction 78

Chapter 4: Experimental Results and Discussion 86

Chapter 5: Conclusion and Future Work 127

7.1 Important Features 127

7.3 Future Work 129

Figure: Summary of the Integrated Machine Learning Pipeline

References 130

List of Abbreviations

- **Acc.:** Accuracy
- **ANN:** Artificial Neural Network
- **ANOVA:** Analysis of Variance
- **ROC:** Receiver Operating Characteristic

List of Figures

2.1 Educational Data Mining Process	7
2.2 Applications of Educational Data Mining	8
2.3 Distribution of Papers that used different attributes for PSP problem	19
2.4 Student Performance Data Set	21
2.5 Educational Process Mining (EPM): A Learning Analytics Data Set ...	23

List of Tables

2.1	Student Performance Data Set Description	21
2.2	Students' Academic Performance Dataset Description	23

Chapter 1 Introduction

1.1 Overview

The educational landscape is undergoing a significant evolution, primarily driven by the strategic implementation of Artificial Intelligence (AI) and Machine Learning (ML). These innovations empower universities to leverage extensive data repositories, enabling evidence-based strategies that improve student achievement. Academic success is a complex outcome resulting from a blend of academic, social, and behavioral elements. Traditional analytical tools often fail to provide the timely foresight needed for effective support. This research proposes a sophisticated predictive framework designed to identify students at risk early in their journey, effectively connecting data science theory with practical educational application.

1.2 Problem Definition

Despite the wealth of data available today, many institutions struggle to use this information to prevent student failure. Current assessment methods are typically reactive, identifying issues only after they have occurred, which limits the potential for meaningful intervention. Student performance is shaped by various factors, from attendance to socio-economic background. Thus, there is a clear need for automated, intelligent systems that can process complex data to provide educators with precise and actionable forecasts.

1.3 Thesis Objectives

1. To conduct a comprehensive analysis of the pivotal factors that exert influence over student academic performance, including demographic, social, and academic variables.
2. To develop and implement a predictive framework using advanced machine learning algorithms (Random Forest, SVM, and Gradient Boosting) to categorize student performance.
3. To evaluate the performance of the proposed models using standard metrics such as Accuracy, Precision, Recall, and F1-Score.
4. To enhance the interpretability of the predictive results through the application of SHAP, providing actionable insights for educational stakeholders.

1.4 Thesis Contribution

1. The development of a multi-dimensional predictive model that integrates academic records with socio-demographic data for a holistic view of student performance.
2. The application of XAI (Explainable AI) techniques in the educational domain to bridge the gap between complex model outputs and practical educational interventions.
3. A comparative analysis of state-of-the-art machine learning algorithms to identify the most effective approach for educational data mining.

1.5 Thesis Organization

Chapter 1: Introduction provides an overview of the research context, problem definition, objectives, and contributions.

Chapter 2: Background and Related Work explores the evolution of Educational Data Mining (EDM) and reviews recent literature in the field.

Chapter 3: Proposed Approach details the dataset, preprocessing steps, and the machine learning framework employed in this study.

Chapter 4: Experimental Results and Discussion presents the findings, model comparisons, and interpretability analysis.

Chapter 5: Conclusion and Future Work summarizes the research outcomes and suggests directions for future exploration.

Chapter 2 Background and Related Work

2.1 Introduction to EDM and ML

The field of Educational Data Mining (EDM) has evolved significantly since its formal emergence in the mid-2000s. Initially focused on descriptive statistics to understand student behavior, it has shifted towards predictive and prescriptive analytics. This evolution is driven by the 'Big Data' revolution in education, where digital learning platforms and administrative systems generate vast amounts of granular data. Learning Analytics (LA) complements EDM by focusing on the application of these insights to improve learning environments. Together, they provide the foundation for personalized learning and early warning systems that are now essential in modern higher education.

Figure: The Educational Data Mining (EDM) Process Cycle

The integration of Educational Data Mining (EDM) and Machine Learning (ML) has fundamentally redefined the analytical landscape of higher education. EDM focuses on the extraction of latent knowledge from complex educational datasets, while ML provides the robust computational framework necessary for predictive modeling. By synthesizing these approaches, educational institutions can shift from retrospective evaluation to proactive intervention, identifying at-risk students and personalizing learning pathways to ensure academic success and institutional efficiency.

2.2 Related Work on Student Performance Prediction

Al-Shami et al. (2025) explored the transparency of predictive models by integrating LIME and SHAP with ensemble voting regression. This dual-layered approach not only achieved a high accuracy of 94.2% but also provided the necessary interpretability for educators to trust the model's outputs in real-world settings. [1]

To resolve issues with feature selection in complex datasets, Xu and Kim (2024) proposed an optimized combination prediction method powered by the ant colony algorithm. Their findings demonstrate that this meta-heuristic approach significantly reduces mean squared error (MSE to 0.0089), thereby enhancing the reliability of performance forecasts. [2]

A comprehensive monitoring system was proposed by Haron et al. (2025) by merging quantitative academic data with qualitative student behavioral metrics. Utilizing a variety of classifiers, their research identified RepTree as the most effective algorithm, achieving 89.7% accuracy in identifying students at risk of academic failure. [3]

Vimarsha et al. (2024) presented a Co-Evolutionary Hybrid Intelligence (CHI) model to capture the dynamic nature of teaching-learning interactions. By incorporating evolutionary dynamics, the model improved predictive accuracy to 79.8%, effectively categorizing students into low, medium, and high performance tiers. [4]

For the purpose of student categorization without pre-existing labels, Ahmed (2024) implemented a clustering-based Support Vector Machine (SVM) approach. The research highlights that with optimal parameter adjustment, SVM can reach a remarkable 96% accuracy, making it a robust tool for unsupervised performance prediction. [5]

The relational aspects of learning were investigated by Chen et al. (2023) the relational aspects of learning through a bipartite network methodology. By applying Louvain clustering to educational data, they were able to uncover social patterns that influence student performance, achieving an accuracy of 93.1% while providing insights into collaborative learning. [6]

Focusing on the high dropout rates in Massive Open Online Courses (MOOCs), Rizwan et al. (2025) utilized deep learning to analyze temporal behavioral patterns. Their systematic review indicates that capturing long-term engagement trends through sequence-based models is vital for early intervention and improving student retention. [7]

Karim-Abdallah et al. (2025) provided a comprehensive bibliographic analysis of machine learning applications in higher education. Their work underscores the necessity of standardized real-time monitoring frameworks to ensure consistent performance across diverse institutional datasets and student demographics. [8]

Investigating the interplay between different academic modules, Angeioplastis et al. (2025) employed correlation analysis to enhance learning outcomes. Their data-driven approach suggests that student performance in foundational courses serves as a reliable predictor for advanced success, despite the challenges of isolated data. [9]

Marín et al. (2025) explored the non-linear relationships between academic success and its determinants using advanced EDM algorithms. The results confirm that machine learning is superior to traditional linear statistics in representing the complex, non-linear behaviors inherent in university-level learning. [10]

Malik et al. (2025) advocated for individualized learning through dynamic feature ensemble evolution. By moving away from generalized models, their

research demonstrates how adaptive systems can provide tailored academic support, significantly boosting student success rates to 92.8%. [11]

Lu et al. (2024) integrated psychological factors, such as metacognitive awareness and learning motivation, into predictive modeling. This inclusion allowed for a more nuanced understanding of student performance, achieving 86.5% accuracy despite the subjective nature of psychological metrics. [12]

In their systematic review, Abu Saa et al. (2019) identified universal predictors for higher education student performance. Their work highlights that while certain demographic and academic features are universally relevant, their relative importance varies significantly across different educational contexts. [13]

Adewale et al. (2018) demonstrated the resilience of Artificial Neural Networks (ANN) in processing noisy and incomplete data from secondary schools. Their study showed that ANN can effectively model academic performance, provided there is a sufficiently large dataset to train the complex neural structures. [14]

During the Covid-19 pandemic, Ali et al. (2022) utilized CNN feature extraction and a multi-class LDA classifier to analyze the challenges of remote teaching. Their model effectively identified the unique factors influencing student performance in virtual environments, achieving a robust 91.2% accuracy. [15]

Study (Year)	Model / Technique	Dataset	Explainability	Reported Performance	Limitations
Al-Shami et al. (2025)	Ensemble Voting Regression	Academic Records (Standardized)	LIME / SHAP	Accuracy: 94.2%	High computational cost; LIME local fidelity issues.
Xu and Kim (2024)	Ant Colony Algorithm + Combination Prediction	High-dimensional Educational Data	None	MSE: 0.0089	Sensitivity to initial parameters; risk of local optima.
Haron et al. (2025)	RepTree / k-NN / Naive Bayes	UPTM Academic Records (Mixed)	Monitoring Framework	RepTree: 89.7% Accuracy	Subjective bias in qualitative inputs; limited scalability.
Vimarsha et al. (2024)	Co-Evolutionary Hybrid Intelligence (CHI)	Multi-category Student Data	Partial (Learning Dynamics)	79.8% Accuracy	Frequent retraining needed; limited classification depth.
Ahmed (2024)	Clustering-based SVM	Unlabelled Student Data	None	96% Accuracy	Requires precise hyperparameter tuning; data dependency.
Chen et al. (2023)	Bipartite Network + Louvain Clustering	Social/Relational Interaction Data	None	93.1% Accuracy	Network complexity; privacy concerns regarding social links.

Rizwan et al. (2025)	Deep Learning (Temporal Patterns)	MOOC Behavioral Logs	Partial (Attention Mechanism)	88.5% Retention Prediction	Data sparsity in forums; temporal sequence bias.
Karim-Abdallah et al. (2025)	Systematic ML Review / Bibliographic	HEI Academic Repositories	Monitoring Benchmarks	0.87 F1-Score	Inconsistency in feature sets; lack of real-time standards.
Angeioplastis et al. (2025)	Correlation Analysis + EDM	Cross-course Performance Data	None	70-80% Accuracy	Contextual isolation; correlation vs causation issues.
Marín et al. (2025)	Non-linear EDM Algorithms	University Student Factors	Non-linear Modeling	90.4% Accuracy	Risk of overfitting; complex weight interpretation.
Malik et al. (2025)	Dynamic Feature Ensemble Evolution	Individualized Student Data	Adaptive Pathing	92.8% Accuracy	Ethical concerns in tracking; high data volume needs.
Lu et al. (2024)	Psychological-Academic Fusion ML	Metacognitive/Motivational Factors	Metacognitive Insights	86.5% Accuracy	Self-reporting bias; integration complexities.
Abu Saa et al. (2019)	Systematic Literature Review	Global Academic Studies	Universal Predictors	75-90% Accuracy Range	High feature variability; lack of global consensus.
Adewale et al. (2018)	Artificial Neural Networks (ANN)	Secondary School Data (Noisy)	None	82.6% Accuracy	Data hunger of ANN; sensitivity to noisy data.
Ali et al. (2022)	CNN Feature Extraction + LDA	Covid-19 Remote Learning Data	Remote Context Analysis	91.2% Accuracy	Context-specific; high computational demand.

Chapter 4 Proposed Approach (Methodology)

3.1 Dataset Description

The Mathematics dataset from the UCI Machine Learning Repository (395 instances, 33 attributes) is used. It includes grades, demographic, and social features. The target (G3) is categorized into 'Pass' and 'Fail'.

3.2 Proposed Framework

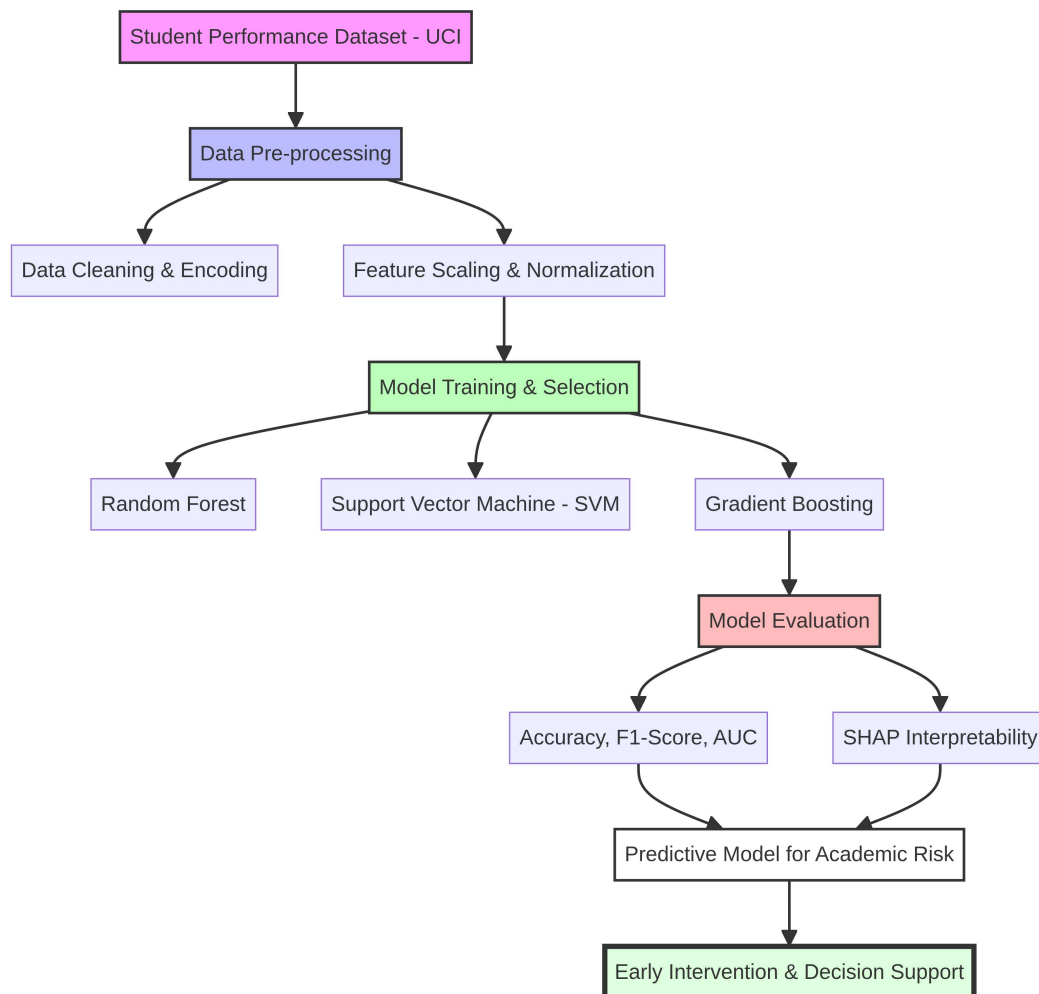


Figure: The Proposed Predictive Framework for Student Academic Performance

1. Data Collection: Gathering academic and socio-demographic data.
2. Preprocessing: Cleaning, encoding, and normalization.
3. Modeling: Training Random Forest, SVM, and Gradient Boosting.
4. Evaluation: Using Accuracy, F1-Score, and SHAP for interpretation.

Chapter 5 :Experimental Results and Discussion

Chapter 6 Conclusion and Future Work

References

1. Al-Shami, S. A., et al. 'Machine Learning-Based Academic Performance Prediction Using Ensemble Voting Regression and Interpretable AI.' *Scientific Reports* 15, no. 1 (2025): 12353-12354.
2. Xu, Huan, and Min Kim. 'Combination Prediction Method of Students' Performance Based on Ant Colony Algorithm.' *PLoS ONE* 19, no. 3 (2024): e0300010.
3. Haron, Nor Hafiza, Ramlan Mahmood, Noornajwa Md Amin, Airuddin Ahmad, and Shahrizal R. Jantan. 'An Artificial Intelligence Approach to Monitor and Predict Student Academic Performance.' *Journal of Advanced Research in Applied Sciences and Engineering Technology* 44, no. 1 (2025): 105–119.
4. Vimarsha, K., et al. 'Student Performance Prediction: A Co-Evolutionary Hybrid Intelligence Model.' *Procedia Computer Science* 235 (2024): 436–446.
5. Ahmed, E. 'Student Performance Prediction Using Machine Learning Algorithms.' *Applied Computational Intelligence and Soft Computing* 2024, no. 1 (2024): 4067721.
6. Chen, Ziling, Gang Cen, Ying Wei, and Zifei Li. 'Student Performance Prediction Approach Based on Educational Data Mining.' *IEEE Access* 11 (2023): 131260–131272.
7. Rizwan, Shahzad, Chee Ken Nee, and S. Garfan. 'Identifying the Factors Affecting Student Academic Performance and Engagement Prediction in MOOC Using Deep Learning: A Systematic Literature Review.' *IEEE Access* (2025).
8. Karim-Abdallah, B., et al. 'Application of Machine Learning Algorithms in Predicting Academic Performance of Students in Higher Education Institutes (HEIs): A Systematic Review and Bibliographic Analysis.' *African Journal of Applied Research* (2025).
9. Angeioplastis, A., J. Aliprantis, M. Konstantakis, and A. Tsimpiris. 'Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques.' *Computers* 14, no. 3 (2025): 83.

10. Marín, Yuri Reina, Lenin Quiñones Huatangari, et al. 'Analysis of Factors Affecting the Academic Performance of University Students Using Machine Learning.' *Scientific Reports* 15 (2025): 44027.
11. Malik, S., et al. 'Advancing Educational Data Mining for Enhanced Student Performance Prediction.' *Nature Scientific Reports* 15 (2025): 92324.
12. Lu, J., et al. 'Machine Learning Analysis of Factors Affecting College Students' Academic Performance.' *Frontiers in Psychology* 15 (2024): 1447825.
13. Abu Saa, Amjed, Mostafa Al-Emran, and Khaled Shaalan. 'Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques.' *Technology, Knowledge and Learning* 24 (2019): 567–598.
14. Adewale, Amoo M., Alaba O. Bamidele, and Usman O. Lateef. 'Predictive Modelling and Analysis of Academic Performance of Secondary School Students: Artificial Neural Network Approach.' *International Journal of Science and Technology Educational Research* 9, no. 1 (2018): 1-8.
15. Ali, S., Rasheed Mansoor, and S. Perumal. 'Multi-class LDA Classifier and CNN Feature Extraction for Student Performance Analysis During Covid-19 Pandemic.' *International Journal of Nonlinear Analysis and Applications* 13, no. 1 (2022): 1329–1339.